METHOD AND APPARATUS USING HARMONIC-MODEL-BASED FRONT END FOR ROBUST SPEECH RECOGNITION

BACKGROUND OF THE INVENTION

5   The present invention relates to noise reduction. In particular, the present invention relates to removing noise from signals used in speech recognition.

A pattern recognition system, such as a 10   speech recognition system, takes an input signal and attempts to decode the signal to find a pattern represented by the signal. For example, in a speech recognition system, a speech signal (often referred to as a test signal) is received by the recognition 15   system and is decoded to identify a string of words represented by the speech signal.

To decode the incoming test signal, most recognition systems utilize one or more models that describe the likelihood that a portion of the test 20   signal represents a particular pattern. Examples of such models include Neural Nets, Dynamic Time Warping, segment models, and Hidden Markov Models.

Before a model can be used to decode an incoming signal, it must be trained. This is 25   typically done by measuring input training signals generated from a known training pattern. For example, in speech recognition, a collection of speech signals is generated by speakers reading from a known text. These speech signals are then used to 30   train the models.

In order for the models to work optimally, the signals used to train the models should be similar to the eventual test signals that are decoded. In particular, the training signals should have the same amount and type of noise as the test signals that are decoded.

Typically, the training signal is collected under "clean" conditions and is considered to be relatively noise free. To achieve this same low level of noise in the test signal, many prior art systems apply noise reduction techniques to the testing data.

One technique for removing noise attempts to model the noise using a set of training signals collected under various noisy conditions. However, such systems are only effective if the noise conditions of the training signals match the noise conditions of the test signals. Because of the large number of possible noises and the seemingly infinite combinations of noises, it is very difficult to build noise models that can handle every test condition.

Another technique for removing noise is to estimate the noise and then subtract it from the noisy speech signal. Typically, such systems estimate the noise from a previous frame of the input signal. As such, if the noise is changing over time, the estimate of the noise for the current frame will be inaccurate.

One system of the prior art for estimating the noise in a speech signal uses the harmonics of

human speech. The harmonics of human speech produce peaks in the frequency spectrum. By identifying nulls between these peaks, these systems identify the spectrum of the noise. This spectrum is then subtracted from the spectrum of the noisy speech signal to provide a clean speech signal.

The harmonics of speech have also been used in speech coding to reduce the amount of data that must be sent when encoding speech for transmission across a digital communication path. Such systems attempt to separate the speech signal into a harmonic component and a random component. Each component is then encoded separately for transmission. One system in particular used a harmonic+noise model in which a sum-of-sinusoids model is fit to the speech signal to perform the decomposition.

In speech coding, the decomposition is done to find a parameterization of the speech signal that accurately represents the input noisy speech signal. The decomposition has no noise-reduction capability and as such has not been useful in speech recognition.

SUMMARY OF THE INVENTION

A system and method are provided that reduce noise in speech signals. The system and method decompose a noisy speech signal into a harmonic component and a residual component. The harmonic component and residual component are then combined as a sum to form a noise-reduced value. In

some embodiments, the sum is a weighted sum where the harmonic component is multiplied by a scaling factor. In some embodiments, the noise-reduced value is used in speech recognition.

5

BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1 is a block diagram of one computing environment in which the present invention may be practiced.

10   FIG. 2 is a block diagram of an alternative computing environment in which the present invention may be practiced.

FIG. 3 is a flow diagram of a method of using a noise reduction system of one embodiment of

15   the present invention.

FIG. 4 is a block diagram of a noise reduction system of one embodiment of the present invention.

FIG. 5 is a block diagram of pattern

20   recognition system with which embodiments of the present invention may be practiced.

DETAILED DESCRIPTION OF ILLUSTRATIVE EMBODIMENTS

FIG. 1 illustrates an example of a suitable computing system environment 100 on which the

25   invention may be implemented. The computing system environment 100 is only one example of a suitable computing environment and is not intended to suggest any limitation as to the scope of use or functionality of the invention. Neither should the

30   computing environment 100 be interpreted as having

any dependency or requirement relating to any one or combination of components illustrated in the exemplary operating environment 100.

5　The invention is operational with numerous other general purpose or special purpose computing system environments or configurations. Examples of well-known computing systems, environments, and/or configurations that may be suitable for use with the invention include, but are not limited to, personal

10　computers, server computers, hand-held or laptop devices, multiprocessor systems, microprocessor-based systems, set top boxes, programmable consumer electronics, network PCs, minicomputers, mainframe computers, telephony systems, distributed computing

15　environments that include any of the above systems or devices, and the like.

The invention may be described in the general context of computer-executable instructions, such as program modules, being executed by a

20　computer. Generally, program modules include routines, programs, objects, components, data structures, etc. that perform particular tasks or implement particular abstract data types. The invention is designed to be practiced in distributed

25　computing environments where tasks are performed by remote processing devices that are linked through a communications network. In a distributed computing environment, program modules are located in both local and remote computer storage media including

30　memory storage devices.

With reference to FIG. 1, an exemplary system for implementing the invention includes a general-purpose computing device in the form of a computer 110. Components of computer 110 may

5 include, but are not limited to, a processing unit 120, a system memory 130, and a system bus 121 that couples various system components including the system memory to the processing unit 120. The system bus 121 may be any of several types of bus structures

10 including a memory bus or memory controller, a peripheral bus, and a local bus using any of a variety of bus architectures. By way of example, and not limitation, such architectures include Industry Standard Architecture (ISA) bus, Micro Channel

15 Architecture (MCA) bus, Enhanced ISA (EISA) bus, Video Electronics Standards Association (VESA) local bus, and Peripheral Component Interconnect (PCI) bus also known as Mezzanine bus.

Computer 110 typically includes a variety

20 of computer readable media. Computer readable media can be any available media that can be accessed by computer 110 and includes both volatile and nonvolatile media, removable and non-removable media. By way of example, and not limitation, computer

25 readable media may comprise computer storage media and communication media. Computer storage media includes both volatile and nonvolatile, removable and non-removable media implemented in any method or technology for storage of information such as

30 computer readable instructions, data structures,

program modules or other data. Computer storage media includes, but is not limited to, RAM, ROM, EEPROM, flash memory or other memory technology, CD-ROM, digital versatile disks (DVD) or other optical

5 disk storage, magnetic cassettes, magnetic tape, magnetic disk storage or other magnetic storage devices, or any other medium which can be used to store the desired information and which can be accessed by computer 110. Communication media

10 typically embodies computer readable instructions, data structures, program modules or other data in a modulated data signal such as a carrier wave or other transport mechanism and includes any information delivery media. The term "modulated data signal"

15 means a signal that has one or more of its characteristics set or changed in such a manner as to encode information in the signal. By way of example, and not limitation, communication media includes wired media such as a wired network or direct-wired

20 connection, and wireless media such as acoustic, RF, infrared and other wireless media. Combinations of any of the above should also be included within the scope of computer readable media.

The system memory 130 includes computer

25 storage media in the form of volatile and/or nonvolatile memory such as read only memory (ROM) 131 and random access memory (RAM) 132. A basic input/output system 133 (BIOS), containing the basic routines that help to transfer information between

30 elements within computer 110, such as during start-

up, is typically stored in ROM 131. RAM 132 typically contains data and/or program modules that are immediately accessible to and/or presently being operated on by processing unit 120. By way of

5 example, and not limitation, FIG. 1 illustrates operating system 134, application programs 135, other program modules 136, and program data 137.

The computer 110 may also include other removable/non-removable volatile/nonvolatile computer

10 storage media. By way of example only, FIG. 1 illustrates a hard disk drive 141 that reads from or writes to non-removable, nonvolatile magnetic media, a magnetic disk drive 151 that reads from or writes to a removable, nonvolatile magnetic disk 152, and an

15 optical disk drive 155 that reads from or writes to a removable, nonvolatile optical disk 156 such as a CD ROM or other optical media. Other removable/non-removable, volatile/nonvolatile computer storage media that can be used in the exemplary operating

20 environment include, but are not limited to, magnetic tape cassettes, flash memory cards, digital versatile disks, digital video tape, solid state RAM, solid state ROM, and the like. The hard disk drive 141 is typically connected to the system bus 121 through a

25 non-removable memory interface such as interface 140, and magnetic disk drive 151 and optical disk drive 155 are typically connected to the system bus 121 by a removable memory interface, such as interface 150.

The drives and their associated computer

30 storage media discussed above and illustrated in FIG.

1, provide storage of computer readable instructions, data structures, program modules and other data for the computer 110. In FIG. 1, for example, hard disk drive 141 is illustrated as storing operating system 5   144, application programs 145, other program modules 146, and program data 147. Note that these components can either be the same as or different from operating system 134, application programs 135, other program modules 136, and program data 137. 10 Operating system 144, application programs 145, other program modules 146, and program data 147 are given different numbers here to illustrate that, at a minimum, they are different copies.

A user may enter commands and information 15 into the computer 110 through input devices such as a keyboard 162, a microphone 163, and a pointing device 161, such as a mouse, trackball or touch pad. Other input devices (not shown) may include a joystick, game pad, satellite dish, scanner, or the like. 20 These and other input devices are often connected to the processing unit 120 through a user input interface 160 that is coupled to the system bus, but may be connected by other interface and bus structures, such as a parallel port, game port or a 25 universal serial bus (USB). A monitor 191 or other type of display device is also connected to the system bus 121 via an interface, such as a video interface 190. In addition to the monitor, computers may also include other peripheral output devices such

as speakers 197 and printer 196, which may be connected through an output peripheral interface 195.

The computer 110 is operated in a networked environment using logical connections to one or more remote computers, such as a remote computer 180. The remote computer 180 may be a personal computer, a hand-held device, a server, a router, a network PC, a peer device or other common network node, and typically includes many or all of the elements described above relative to the computer 110. The logical connections depicted in FIG. 1 include a local area network (LAN) 171 and a wide area network (WAN) 173, but may also include other networks. Such networking environments are commonplace in offices, enterprise-wide computer networks, intranets and the Internet.

When used in a LAN networking environment, the computer 110 is connected to the LAN 171 through a network interface or adapter 170. When used in a WAN networking environment, the computer 110 typically includes a modem 172 or other means for establishing communications over the WAN 173, such as the Internet. The modem 172, which may be internal or external, may be connected to the system bus 121 via the user input interface 160, or other appropriate mechanism. In a networked environment, program modules depicted relative to the computer 110, or portions thereof, may be stored in the remote memory storage device. By way of example, and not limitation, FIG. 1 illustrates remote application

programs 185 as residing on remote computer 180. It will be appreciated that the network connections shown are exemplary and other means of establishing a communications link between the computers may be

5 used.

FIG. 2 is a block diagram of a mobile device 200, which is an exemplary computing environment. Mobile device 200 includes a microprocessor 202, memory 204, input/output (I/O)

10 components 206, and a communication interface 208 for communicating with remote computers or other mobile devices. In one embodiment, the afore-mentioned components are coupled for communication with one another over a suitable bus 210.

15 Memory 204 is implemented as non-volatile electronic memory such as random access memory (RAM) with a battery back-up module (not shown) such that information stored in memory 204 is not lost when the general power to mobile device 200 is shut down. A

20 portion of memory 204 is preferably allocated as addressable memory for program execution, while another portion of memory 204 is preferably used for storage, such as to simulate storage on a disk drive.

Memory 204 includes an operating system

25 212, application programs 214 as well as an object store 216. During operation, operating system 212 is preferably executed by processor 202 from memory 204. Operating system 212, in one preferred embodiment, is a WINDOWS® CE brand operating system commercially

30 available from Microsoft Corporation. Operating

system 212 is preferably designed for mobile devices, and implements database features that can be utilized by applications 214 through a set of exposed application programming interfaces and methods. The objects in object store 216 are maintained by applications 214 and operating system 212, at least partially in response to calls to the exposed application programming interfaces and methods.

Communication interface 208 represents numerous devices and technologies that allow mobile device 200 to send and receive information. The devices include wired and wireless modems, satellite receivers and broadcast tuners to name a few. Mobile device 200 can also be directly connected to a computer to exchange data therewith. In such cases, communication interface 208 can be an infrared transceiver or a serial or parallel communication connection, all of which are capable of transmitting streaming information.

Input/output components 206 include a variety of input devices such as a touch-sensitive screen, buttons, rollers, and a microphone as well as a variety of output devices including an audio generator, a vibrating device, and a display. The devices listed above are by way of example and need not all be present on mobile device 200. In addition, other input/output devices may be attached to or found with mobile device 200 within the scope of the present invention.

Under one aspect of the present invention, a system and method are provided that reduce noise in a speech signal by decomposing the speech signal into a harmonic component and a random component and then

5   taking a weighted sum of the harmonic component and the random component to form a noise-reduced feature vector representing a noise-reduced speech signal. The noise-reduced feature vector can then be used as input to a speech decoder.

10   A method and apparatus for forming noise-reduced feature vectors are shown in the flow diagram of FIG. 3 and the block diagram of FIG. 4, respectively.

At step 300, a noisy speech signal is

15   collected and converted into digital samples. To do this, a microphone 404 of FIG. 4, converts audio waves from a speaker 400 and one or more additive noise sources 402 into electrical signals. The electrical signals are then sampled by an analog-to-

20   digital converter 406 to generate a sequence of digital values. In one embodiment, A-to-D converter 406 samples the analog signal at 16 kHz and 16 bits per sample, thereby creating 32 kilobytes of speech data per second. At step 302, the digital samples

25   are grouped into frames by a frame constructor 408. Under one embodiment, frame constructor 408 creates a new frame every 10 milliseconds that includes 25 milliseconds worth of data.

The samples of each frame are provided to a

30   harmonic decompose unit 410, which uses the samples

at step 304 to decompose the signal associated with the frame into a harmonic component and a random component. Thus, the noisy signal is represented as:

$$y = y_h + y_r \qquad \text{EQ. 1}$$

where y is the noisy signal, y_h is the harmonic component, and y_r is the random component.

Under one embodiment, the harmonic component is modeled as a sum of harmonically-related sinusoids such that:

$$y_h = \sum_{k=1}^{K} a_k \cos(k\omega_0 t) + b_k \sin(k\omega_0 t) \qquad \text{EQ. 2}$$

where ω_0 is the fundamental or pitch frequency and K is the total number of harmonics in the signal.

Thus, to identify the harmonic component, an estimate of the pitch frequency and the amplitude parameters $\{a_1 a_2 \ldots a_k b_1 b_2 \ldots b_k\}$ must be determined.

An estimate for the pitch frequency can be determined using any number of available pitch tracking systems. Under many of these systems, candidate pitches are used to identify possible spacing between the centers of segments of the speech signal. For each candidate pitch, a correlation is determined between successive segments of speech. In general, the candidate pitch that provides the best correlation will be the pitch frequency of the frame. In some systems, additional information is used to refine the pitch selection such as the energy of the signal and/or an expected pitch track.

Given an estimate of the pitch, a least-squares solution for the amplitude parameters can be determined. To do so, equation 2 is rewritten as:

$$\mathbf{y} = \mathbf{Ab} \qquad\qquad \text{EQ. } 3$$

where $\mathbf{y}$ is a vector of N samples of the noisy speech signal, $\mathbf{A}$ is an $N \times 2K$ matrix given by:

$$\mathbf{A} = [\mathbf{A}_{\cos}\mathbf{A}_{\sin}] \qquad\qquad \text{EQ. } 4$$

with elements

$$\mathbf{A}_{\cos}(k,t) = \cos(k\omega_0 t) \qquad\qquad \mathbf{A}_{\sin}(k,t) = \sin(k\omega_0 t) \qquad \text{EQ. } 5$$

and b is a $2K \times 1$ vector given by:

$$\mathbf{b}^T = [a_1 a_2 \ldots a_k b_1 b_2 \ldots b_k] \qquad\qquad \text{EQ. } 6$$

Then, the least-squares solution for the amplitude coefficients is:

$$\hat{\mathbf{b}} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{y} \qquad\qquad \text{EQ. } 7$$

Using $\hat{\mathbf{b}}$, an estimate for the harmonic component of the noisy speech signal can be determined as:

$$\mathbf{y}_h = \mathbf{A}\hat{\mathbf{b}} \qquad\qquad \text{EQ. } 8$$

An estimate of the random component is then calculated as:

$$\mathbf{y}_r = \mathbf{y} - \mathbf{y}_h \qquad\qquad \text{EQ. } 9$$

Thus, using equations 7-9 above, harmonic decompose unit 410 is able to produce a vector of harmonic component samples 412, $\mathbf{y}_h$, and a vector of random component samples 414, $\mathbf{y}_r$.

After the samples of the frame have been decomposed into harmonic and random samples, a

scaling parameter or weight is determined for the harmonic component at step 306. This scaling parameter is used as part of a calculation of a noise-reduced speech signal as discussed further

5   below. Under one embodiment, the scaling parameter is calculated as:

$$\alpha_h = \frac{\sum_i y_h(i)^2}{\sum_i y(i)^2} \qquad \text{EQ. 10}$$

where α_h is the scaling parameter, $y_h(i)$ is the ith sample in the vector of harmonic component samples $\mathbf{y}_h$

10 and $y(i)$ is the ith sample of the noisy speech signal for this frame. In Equation 10, the numerator is the sum of the energy of each sample of the harmonic component and the denominator is the sum of the energy of each sample of the noisy speech signal.

15 Thus, the scaling parameter is the ratio of the harmonic energy of the frame to the total energy of the frame.

In alternative embodiments, the scaling parameter is set using a probabilistic voiced-

20 unvoiced detection unit. Such units provide the probability that a particular frame of speech is voiced, meaning that the vocal cords resonate during the frame, rather than unvoiced. The probability that the frame is from a voiced region of speech can

25 be used directly as the scaling parameter.

After the scaling parameter has been determined or while it is being determined, the Mel spectra for the vector of harmonic component samples

and the vector of random component samples are determined at step 308. This involves passing each vector of samples through a Discrete Fourier Transform (DFT) 418 to produce a vector of harmonic

5   component frequency values 422 and a vector of random component frequency values 420. The power spectra represented by the vectors of frequency values are then smoothed by a Mel weighting unit 424 using a series of triangular weighting functions applied

10   along the Mel scale. This results in a harmonic component Mel spectral vector 428, Y_h, and a random component Mel spectral vector 426, Y_r.

   At step 310, the Mel spectra for the harmonic component and the random component are

15   combined as a weighted sum to form an estimate of a noise-reduced Mel spectrum. This step is performed by weighted sum calculator 430 using the scaling factor determined above in the following equation:

$$\hat{X}(t) = \alpha_h(t)Y_h(t) + \alpha_r Y_r(t) \qquad \text{EQ. 11}$$

20   where $\hat{X}(t)$ is the estimate of the noise-reduced Mel spectrum, $Y_h(t)$ is the harmonic component Mel spectrum, $Y_r(t)$ is the random component Mel spectrum, $\alpha_h(t)$ is the scaling factor determined above, α_r is a fixed scaling factor for the random component that in

25   one embodiment is set equal to .1, and the time index t is used to emphasize that the scaling factor for the harmonic component is determined for each frame while the scaling factor for the random component

remains fixed. Note that in other embodiments, the scaling factor for the random component may be determined for each frame.

After the noise-reduced Mel spectrum has 5 been calculated at step 310, the log 432 of the Mel spectrum is determined and then is applied to a Discrete Cosine Transform 434 at step 312. This produces a Mel Frequency Cepstral Coefficient (MFCC) feature vector 436 that represents a noise-reduced 10 speech signal.

A separate noise-reduced MFCC feature vector is produced for each frame of the noisy signal. These feature vectors may be used for any desired purpose including speech coding and speech 15 recognition.

Under one embodiment, the MFCC feature vectors for the noise-reduced speech are used directly in a speech recognition system as shown in FIG. 5.

20 If the input signal is a training signal, the series of MFCC feature vectors 436 for the noise-reduced speech is provided to a trainer 500, which uses the MFCC feature vectors and a training text 502 to train an acoustic model 504. Techniques for 25 training such models are known in the art and a description of them is not required for an understanding of the present invention.

If the input signal is a test signal, the MFCC speech feature vectors are provided to a decoder 30 506, which identifies a most likely sequence of words

based on the stream of feature vectors, a lexicon 508, a language model 510, and the acoustic model 504. The particular method used for decoding is not important to the present invention and any of several 5   known methods for decoding may be used.

Although the invention has been described above in connection with Mel Frequency Cepstral Coefficients, other types of features may be used in the weighted harmonic and noise model of Equation 11.

10   Although the present invention has been described with reference to particular embodiments, workers skilled in the art will recognize that changes may be made in form and detail without departing from the spirit and scope of the invention.